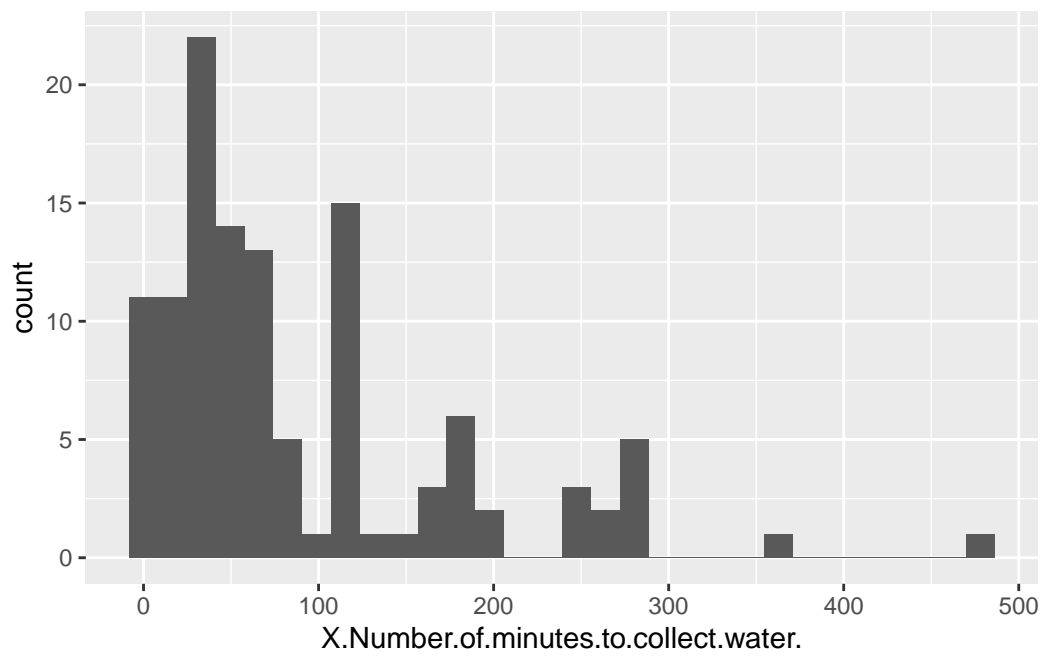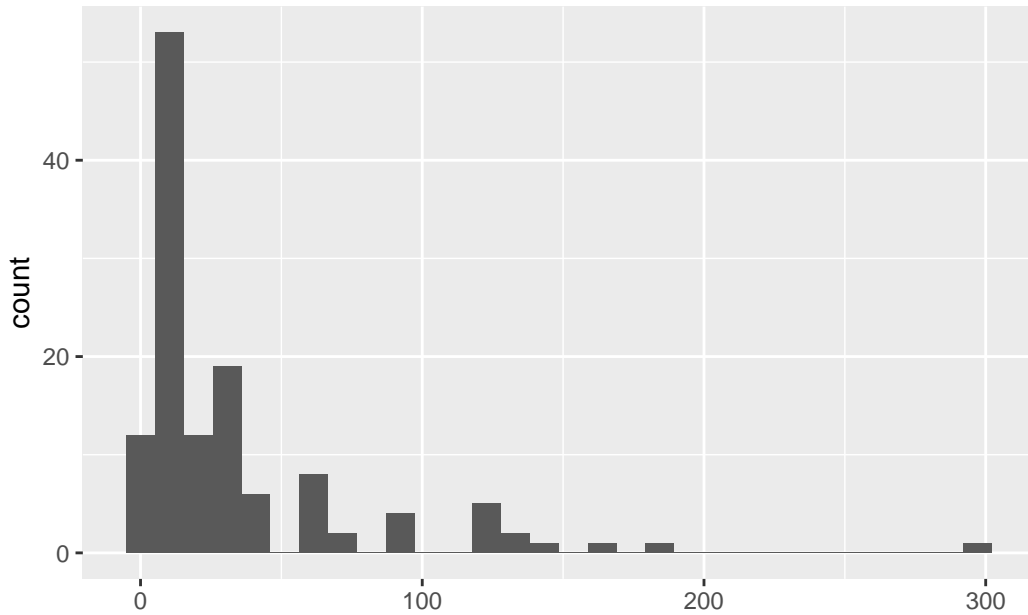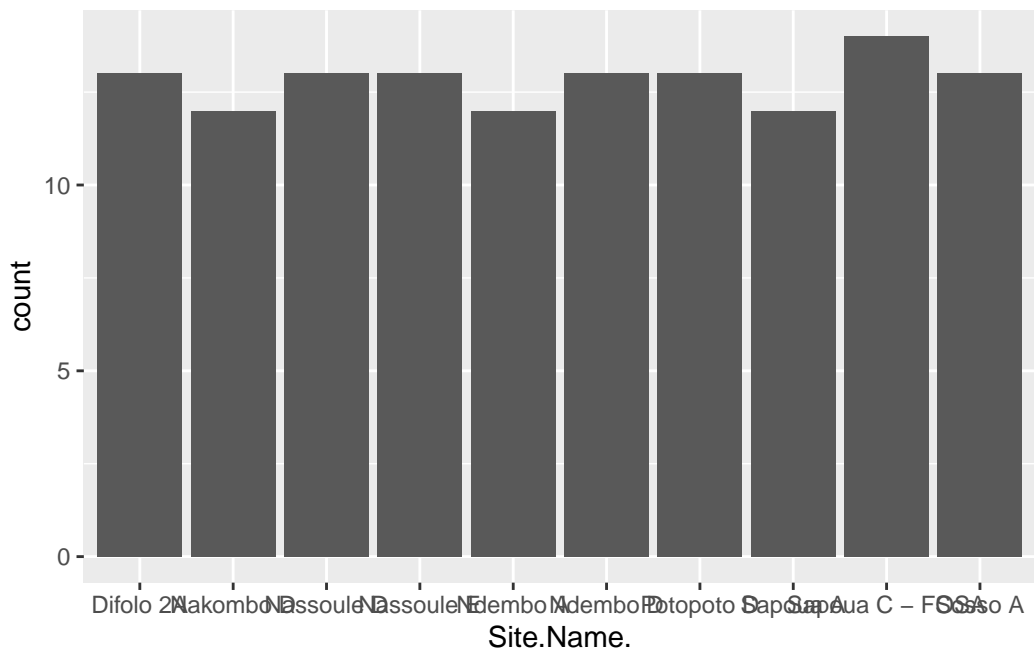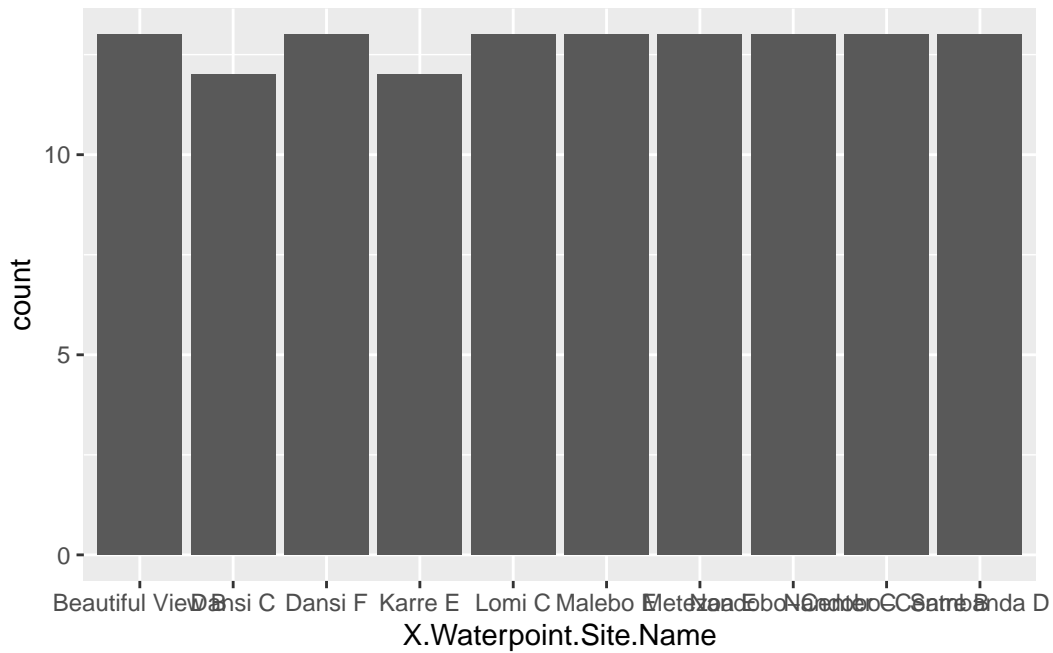# Water for Good

Report

Uni-variate EDA

utes.it.takes.to.walk.from.the.house.to.the.water.point..collect.the.water..including.w

```
# A tibble: 1 x 8
  n_missing numeric.mean numeric.sd numeric.p0 numeric.p25 numeric.p50
      <int>        <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
1        11         88.4       88.0          2          30          60
  numeric.p75 numeric.p100
        <dbl>        <dbl>
1         120          480


# A tibble: 1 x 8
  n_missing numeric.mean numeric.sd numeric.p0 numeric.p25 numeric.p50
      <int>        <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
1         1         34.7       43.1          3          10          15
  numeric.p75 numeric.p100
        <dbl>        <dbl>
1          35          300
```

Both datasets are clearly right skewed therefore using the median amount of time for water collection drastically decreased from the baseline to the recent project survey. The median amount of time for water collection decreased from about 60 to 15 minutes. With the ranges also being drastically smaller for the minutes for the post survey.

2

Based of the both the graphs, there is no overlap in the waterpoint site names. Having a project and baseline, wondering if there is any other data that would help keep one dataset as a base for comparison.
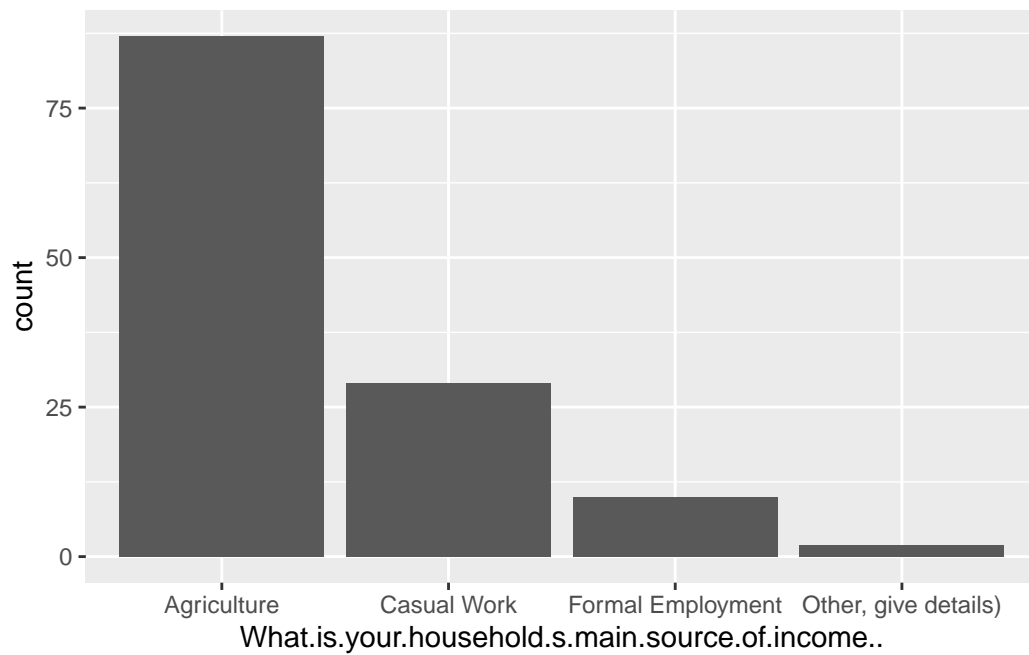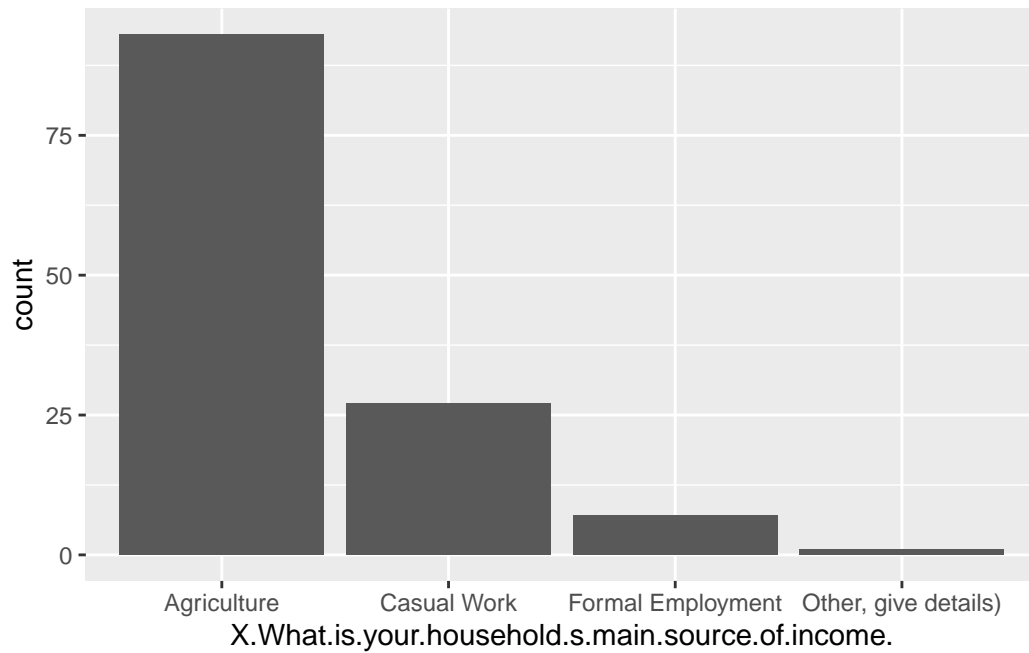
Both have similar household income levels, this could be used a baseline of comparison rather than the waterpoint site name.
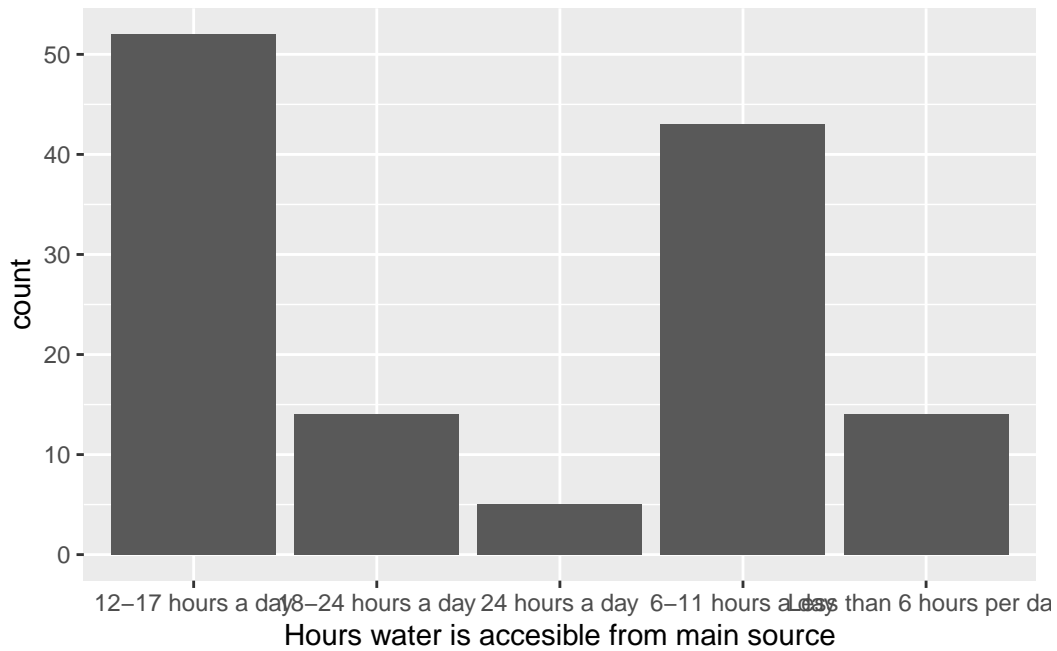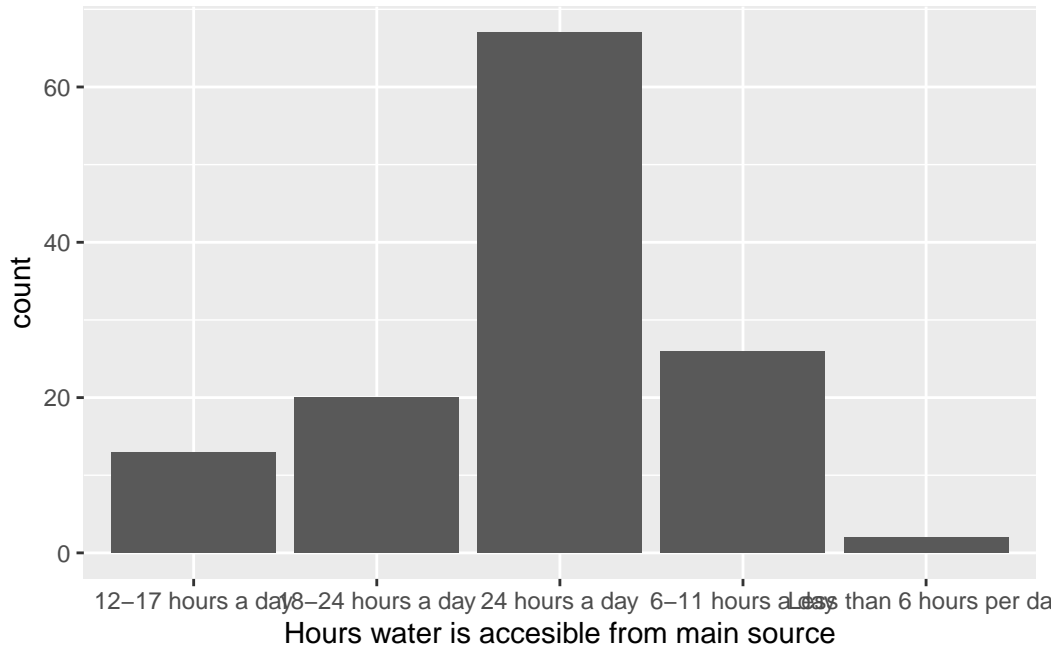
Based on both the graphs, there is a increase in the "12-17 hours a day" water availability post the baseline survey. However, an interesting thing that the 24 hours a day drastically decreased from around 65 to 5 wells available.

Bi-variate EDA
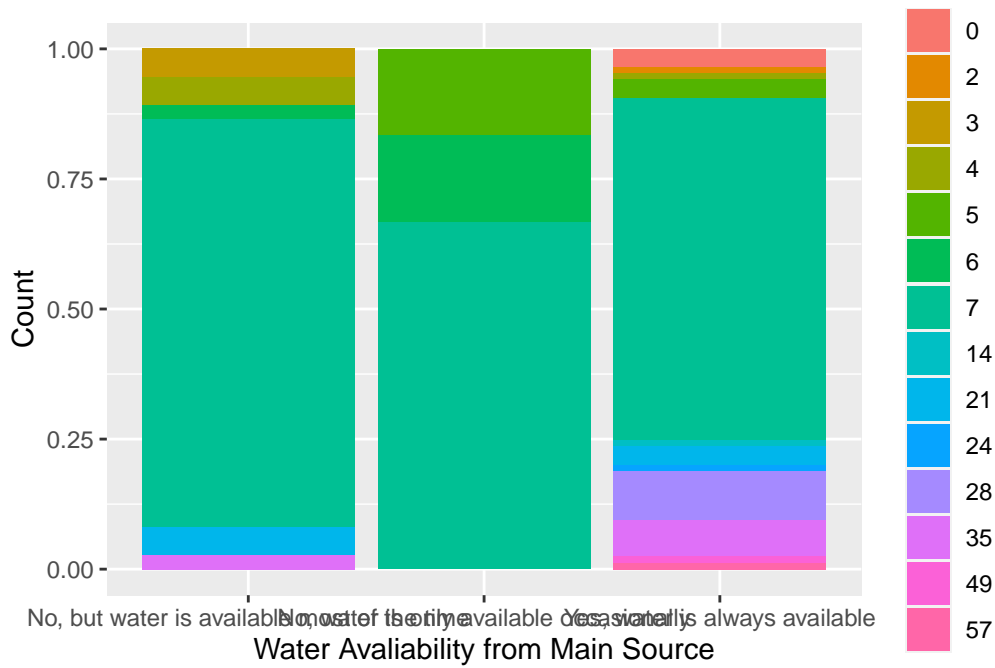
For the instances that water is always available, there seems to be a higher percentage of more trips taken to the main water source. One thing that did stand out to me was the response of "No, but water is available most of the time" had the most amount of "medium" trips taken (around 5-6 trips). Assuming that they had bigger containers to store and carry water, therefore needing less trips. The people who have constant water access, have smaller containers and therefore make more trips in order to obtain water from their main water source.
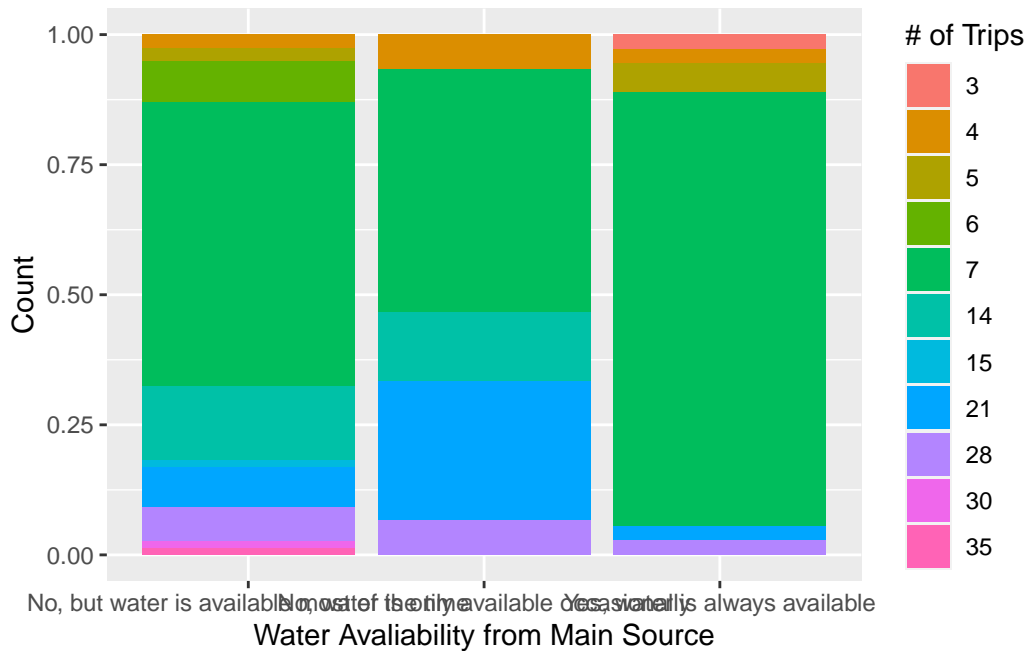
We see a drastic difference in the "Yes, water is always available" column as the number of trips frequented dropped from a high of 24-57 trips to a majority of only 6-7 trips. Assuming the project survey was done months after the wells were drilled, we could see that it helped in reducing the trips made by local residents.

WEEK 2

Linear Regression and Hypothesis Testing

term	estimate	std.error	statistic	p.value
(Intercept)	13.546	1.267	10.689	0.000
Is.water.always.available.from.your.main.water.source..	-2.798	0.964	-2.901	0.004

```
# A tibble: 2 x 2
  term                                              p_value
  <chr>                                               <dbl>
1 Is.water.always.available.from.your.main.water.source..   0.006
2 intercept                                           0.006
```

$Trips = 13.546 - 2.798 * Water\_availability$

Where

No, water is only available occasionally $= 0$, No, but water is available most of the time $= 1$

Yes, water is always available $= 2$

Since p-value is lower than 0.05, we can reject the null hypothesis and conclude that there is a statistically linear relationship between water availability from the waterpoint and number of trips taken. Therefore, whenever there are more water availability from the waterpoint site, the less number of trips taken.

Summary Statistics

Counts by Gender

Interviwee Gender for Post Survey
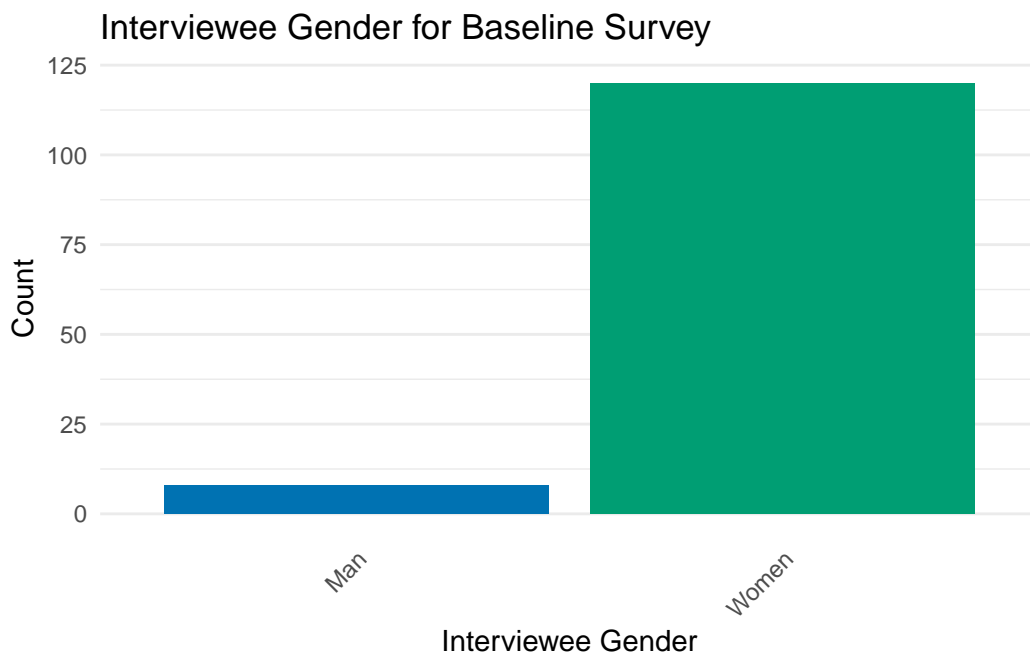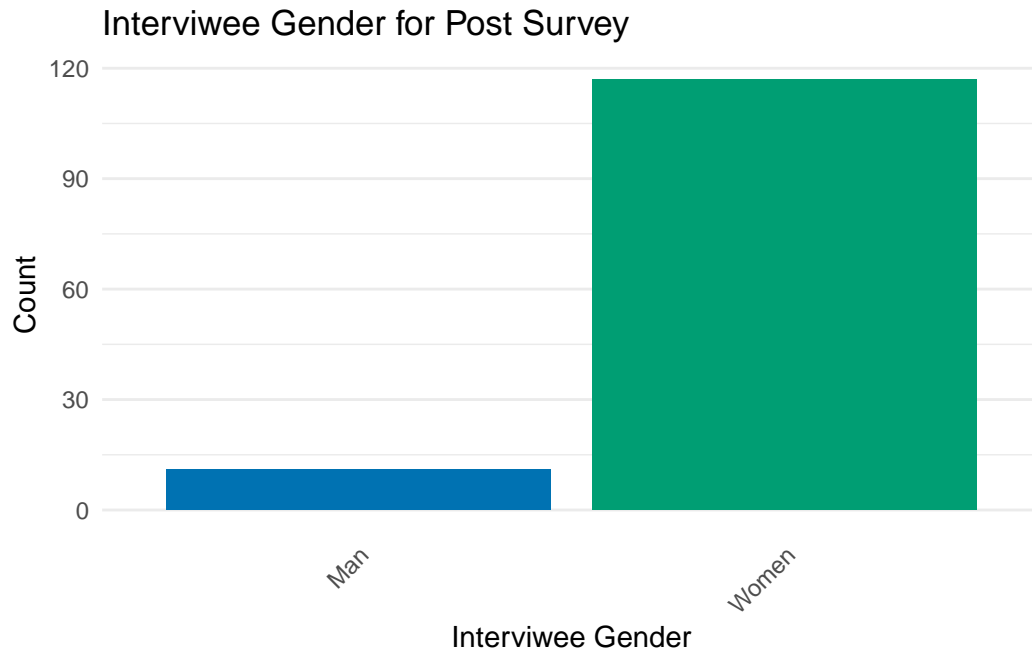
Both baseline and project survey results have similar demographics.

Counts by Education Level

Highest Education for Baseline Survey

Highest Education for Post Survey


```
# A tibble: 4 x 2
  X.What.is.the.highest.level.of.education.obtained.by.the.head.of.the.h~1 count
  <chr>                                                                    <int>
1 None                                                                        18
2 Post-Secondary/University                                                    4
3 Primary                                                                     70
4 Secondary                                                                   36
# i abbreviated name:
#   1: X.What.is.the.highest.level.of.education.obtained.by.the.head.of.the.household.
```
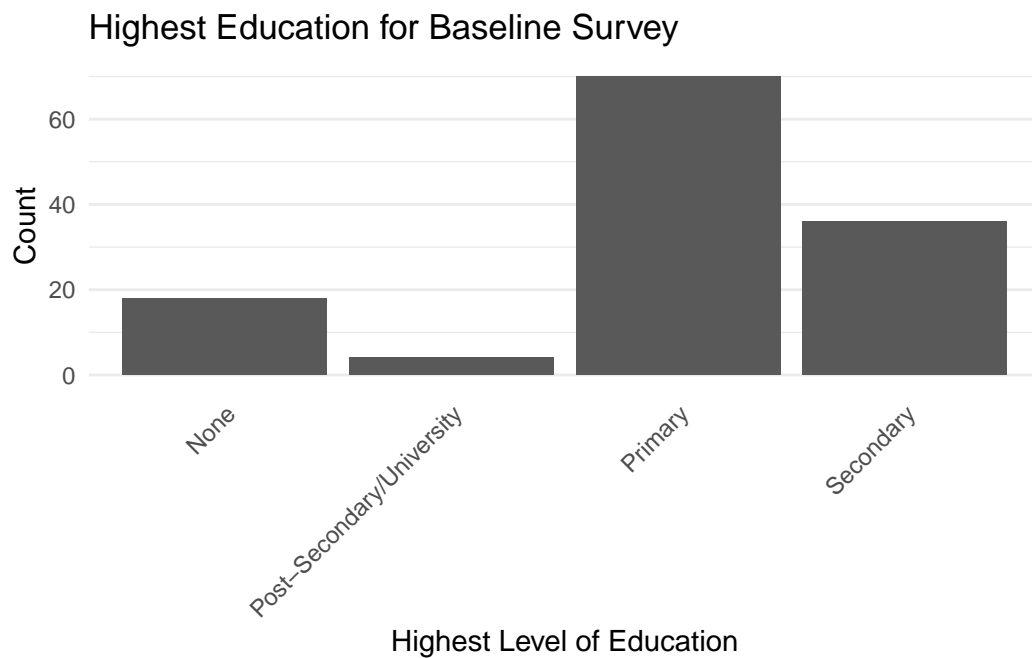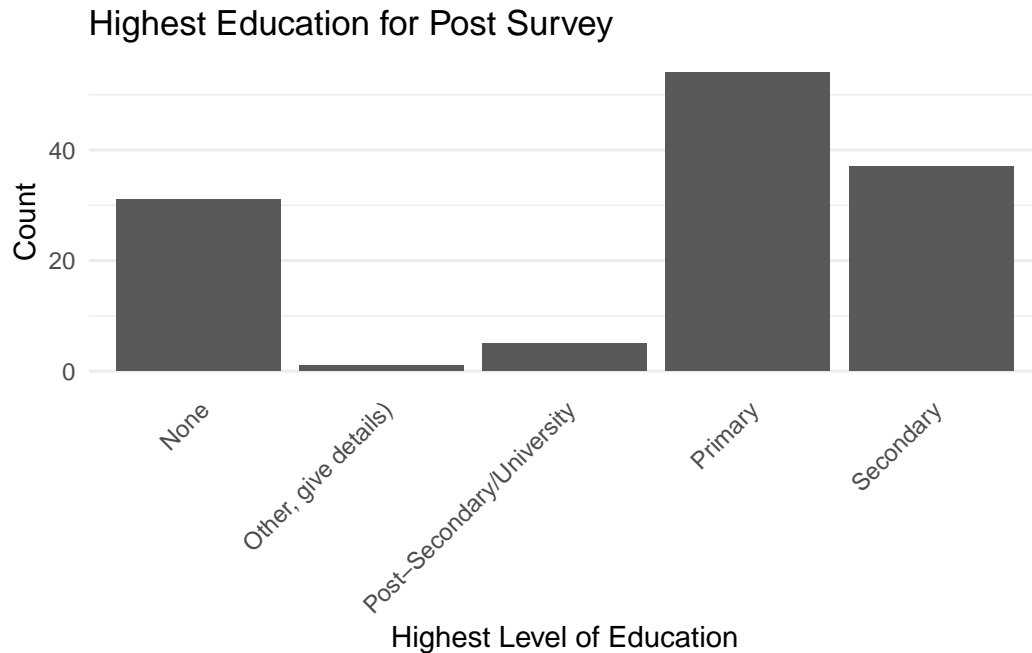
```
# A tibble: 5 x 2
  What.is.the.highest.level.of.education.attained.by.the.head.of.the.hou~1 count
  <chr>                                                                    <int>
1 None                                                                        31
2 Other, give details)                                                         1
3 Post-Secondary/University                                                    5
4 Primary                                                                     54
5 Secondary                                                                   37
# i abbreviated name:
#   1: What.is.the.highest.level.of.education.attained.by.the.head.of.the.household..
```

Both surveys had the same number of participants, 128. Baseline Survey had more people who

completed primary school, but the rest of the distribution regarding secondary, post secondary and other remained the same.

Counts by household income

Income distribution in Baseline Survey

Income distribution in Post Survey
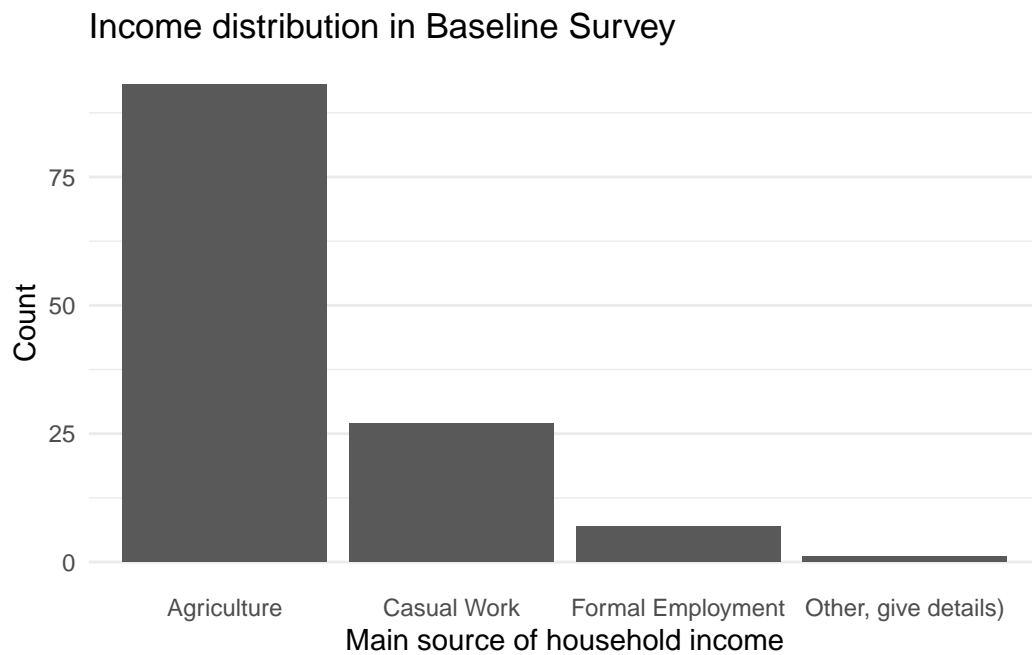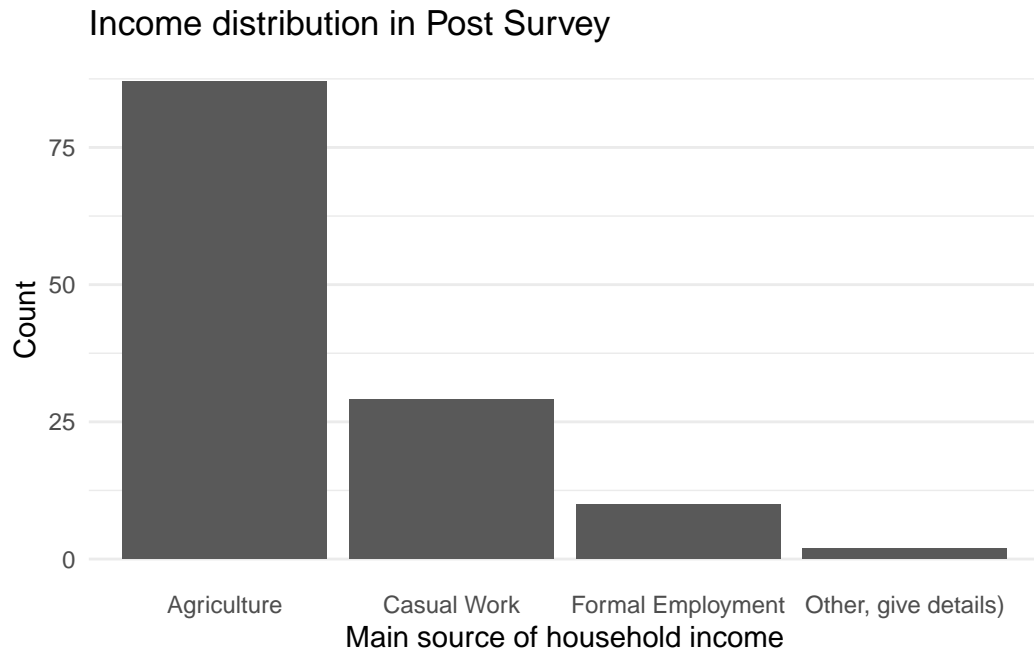
Similar distributions in Income distribution in both baseline and post survey.

Average number of people living in household

Number of people living in house for Baseline Survey

Number of people living in house for Post Survey
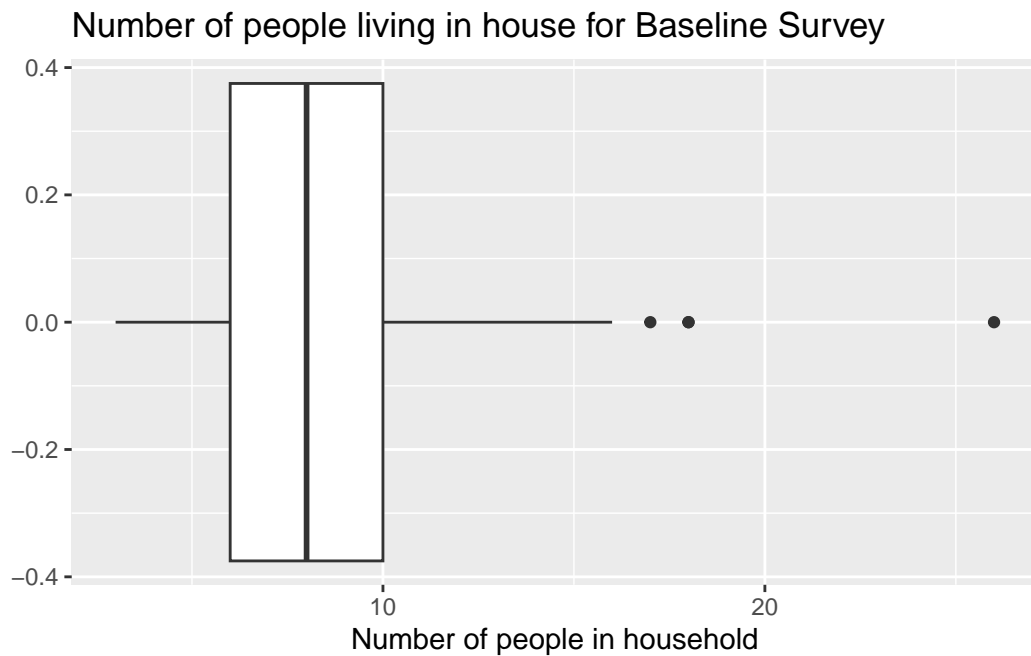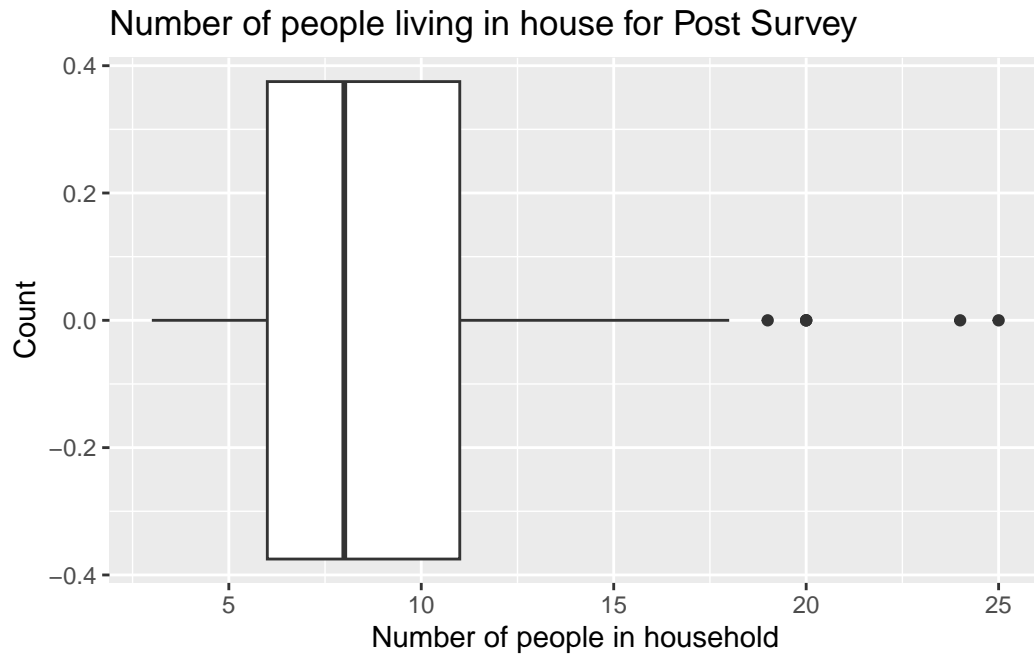

```
# A tibble: 17 x 2
   X.How.many.people.live.in.your.household.in.total..including.yourself. count
                                                                    <int> <int>
 1                                                                      3     3
 2                                                                      4     5
 3                                                                      5    14
 4                                                                      6    19
 5                                                                      7    21
 6                                                                      8    13
 7                                                                      9    12
 8                                                                     10    10
 9                                                                     11     8
10                                                                     12     9
11                                                                     13     2
12                                                                     14     5
13                                                                     15     2
14                                                                     16     1
15                                                                     17     1
16                                                                     18     2
17                                                                     26     1

# A tibble: 19 x 2
```

```
     How.many.people.live.in.your.household.in.total..Ie.the.number.of.peo~1 count
                                                                       <int> <int>
 1                                                                         3     5
 2                                                                         4    10
 3                                                                         5    12
 4                                                                         6    16
 5                                                                         7    17
 6                                                                         8    13
 7                                                                         9     6
 8                                                                        10    15
 9                                                                        11     3
10                                                                        12     8
11                                                                        13     4
12                                                                        14     1
13                                                                        15     8
14                                                                        17     3
15                                                                        18     1
16                                                                        19     1
17                                                                        20     3
18                                                                        24     1
19                                                                        25     1
# i abbreviated name:
#   1: How.many.people.live.in.your.household.in.total..Ie.the.number.of.people.living.under
```
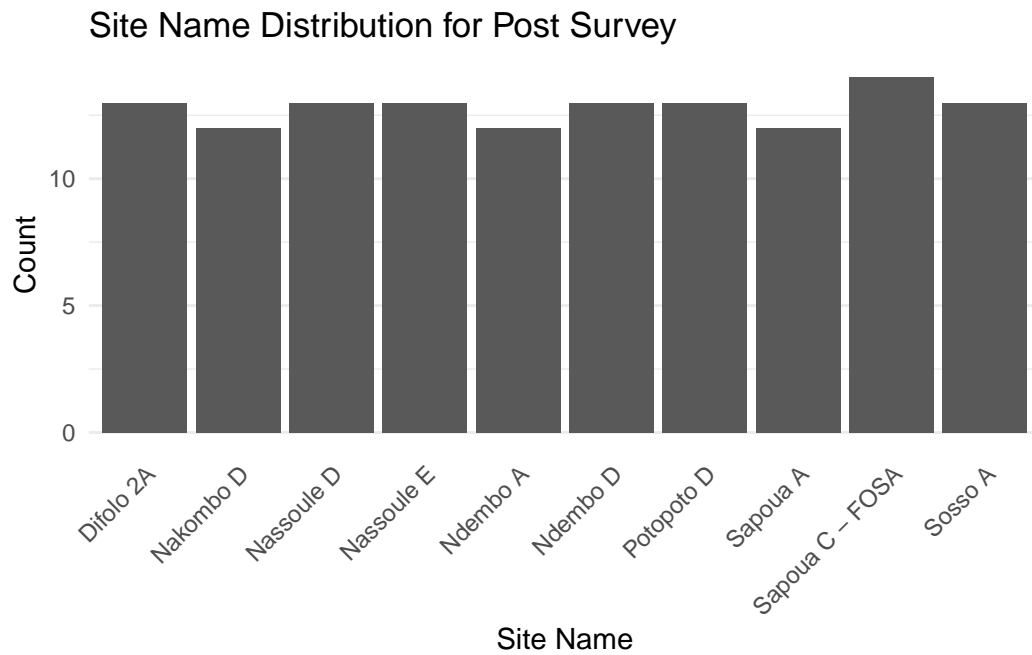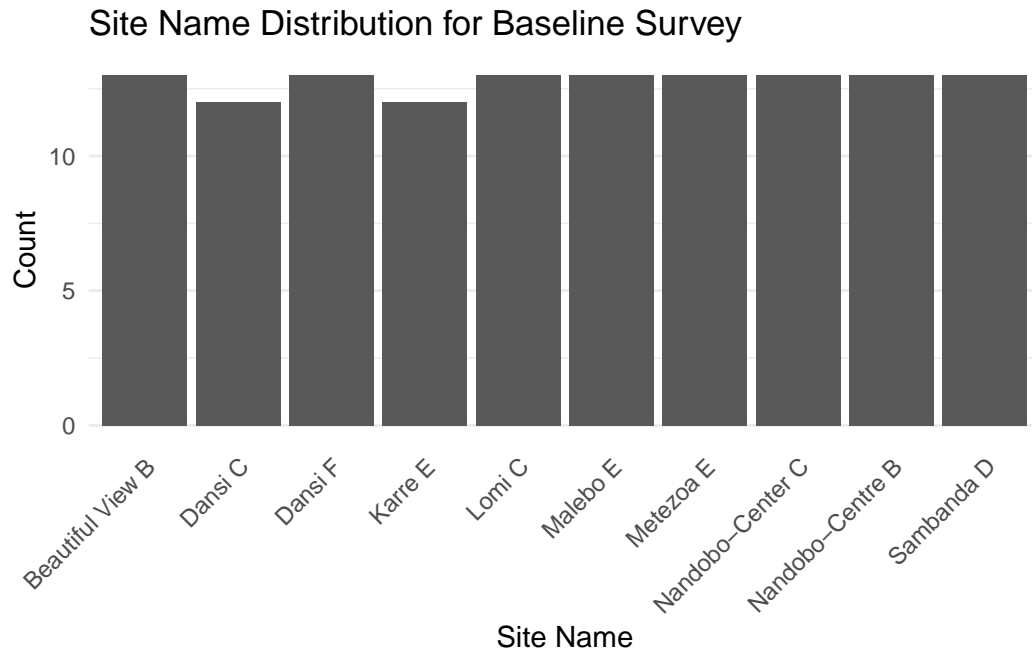
Similar distribution for both baseline and project.

Waterpoint Site Name

Site Name Distribution for Baseline Survey

Site Name Distribution for Post Survey

We can clearly see that baseline and project had different waterpoint site name. Difference between them suggests possibly not the best to compare rather create summary statistics for both surveys individually.
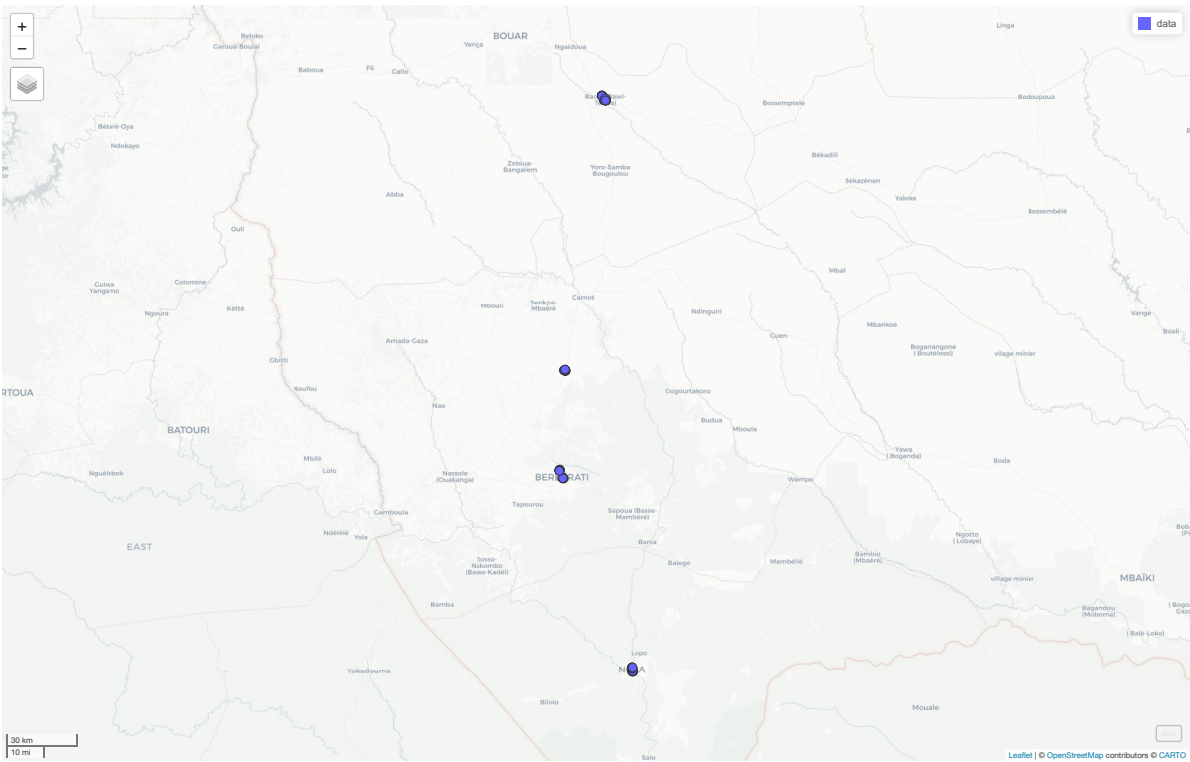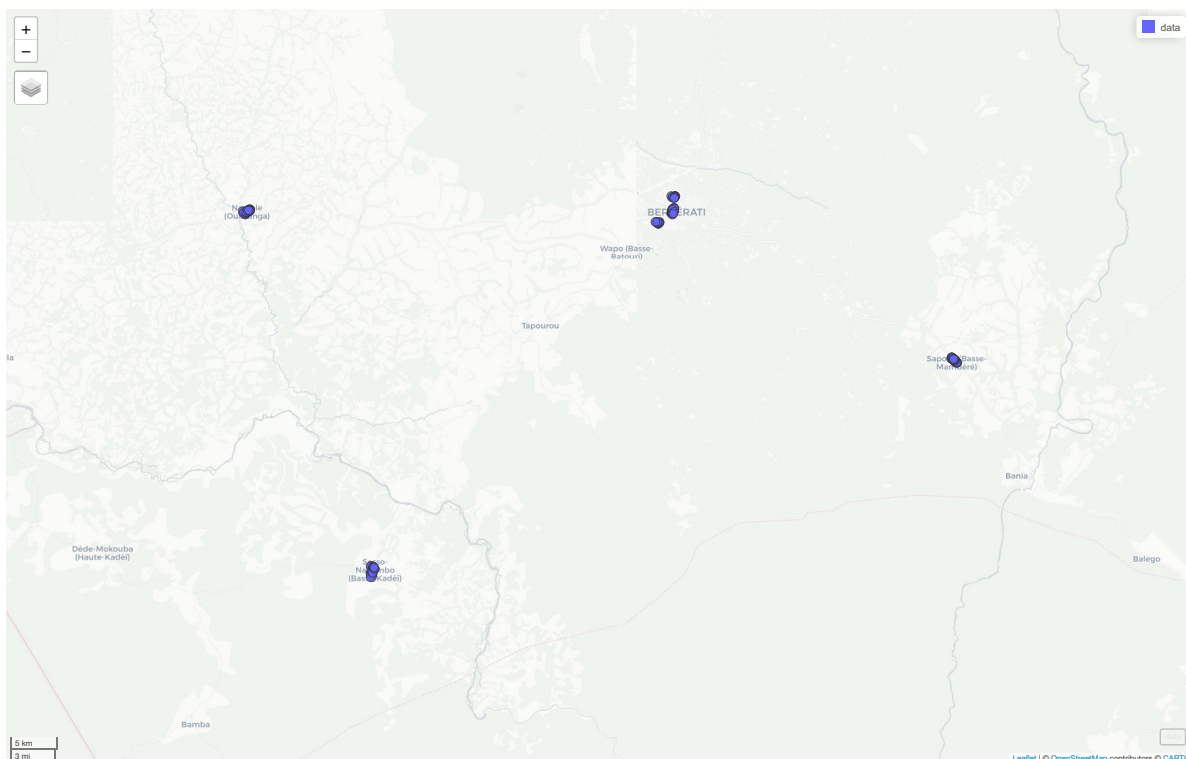
Subjects within Prefecture

Prefecture Distribution for Baseline Survey

Prefecture Distribution for Post Survey
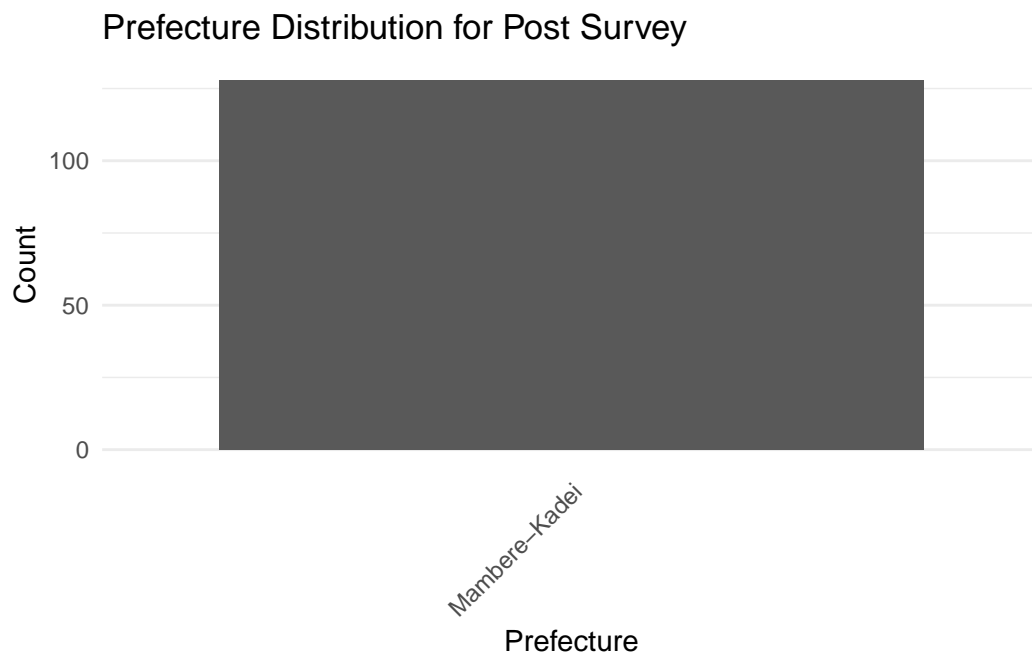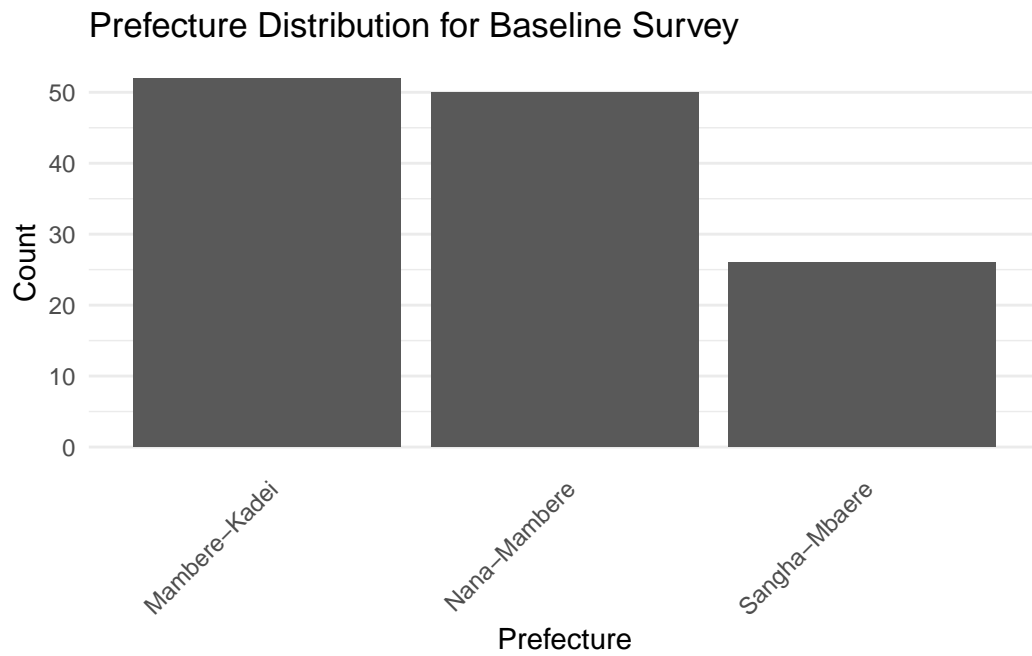
This makes it more clear as the post survey only included residents from the Mambere-Kadei prefecture, which could be the reason why the waterpoint site wells differed from the baseline
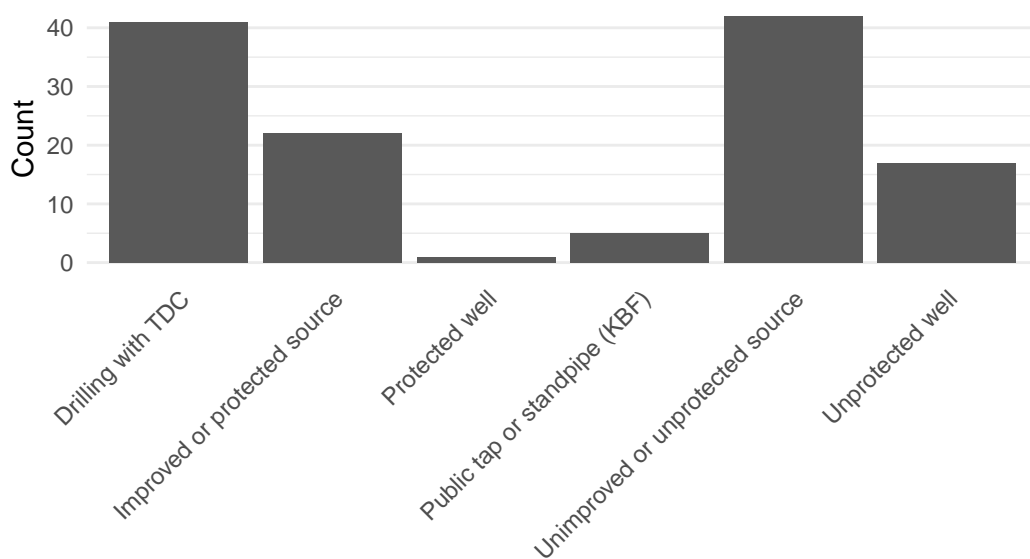
to post.

Map

Categorical answers to survey by Waterpoint site

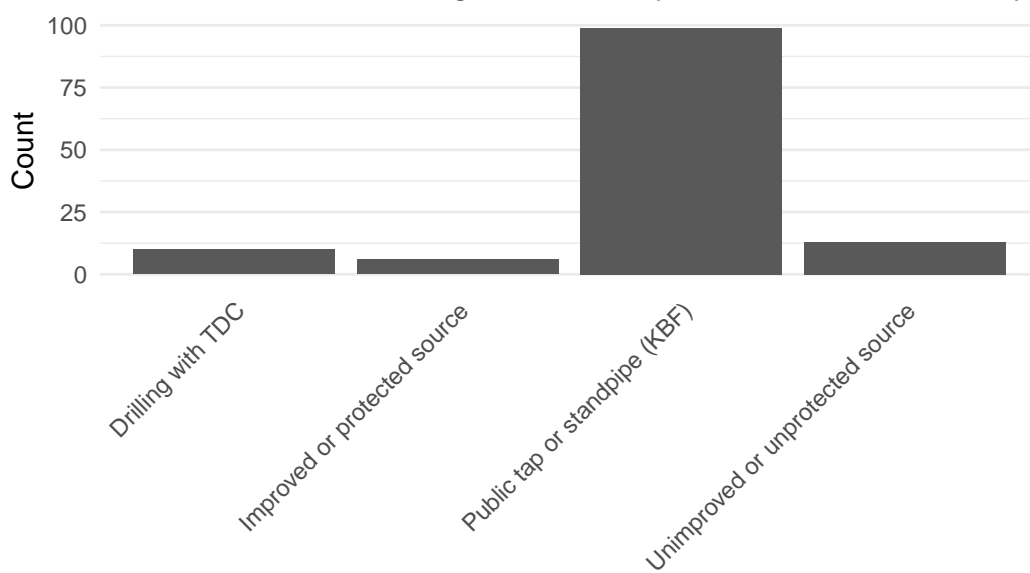Main Source of Drinking Water in Dry Season in Baseline Surve
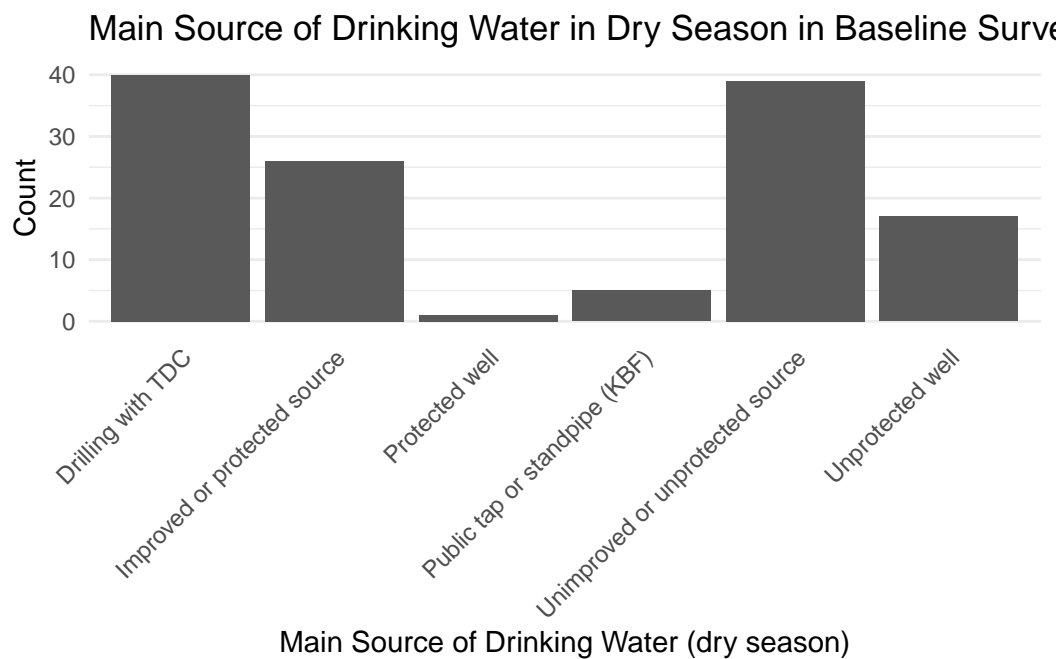
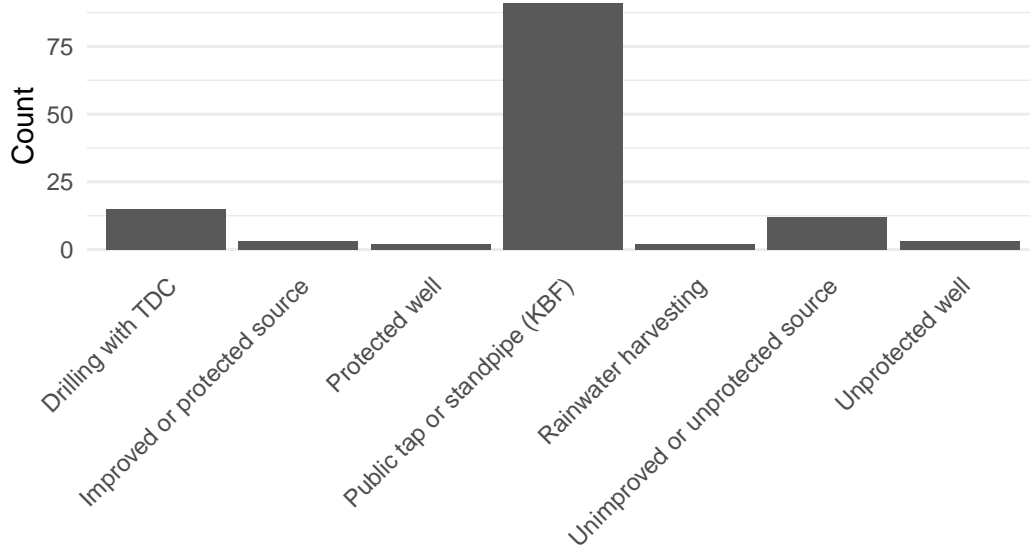## Main Source of Drinking Water in Wet Season in Baseline Surv

Main Source of Drinking Water (wet season)

Main Source of Drinking Water in Dry Season in Post Survey

Main Source of Drinking Water (dry season)

Main Source of Drinking Water in Wet Season in Post Survey



```
# A tibble: 10 x 2
   X.Waterpoint.Site.Name count
   <chr>                  <int>
 1 Beautiful View B          13
 2 Dansi C                   12
 3 Dansi F                   13
 4 Karre E                   12
 5 Lomi C                    13
 6 Malebo E                  13
 7 Metezoa E                 13
 8 Nandobo-Center C          13
 9 Nandobo-Centre B          13
10 Sambanda D                13

# A tibble: 10 x 2
   Site.Name.    count
   <chr>         <int>
 1 Difolo 2A        13
 2 Nakombo D        12
 3 Nassoule D       13
 4 Nassoule E       13
 5 Ndembo A         12
```
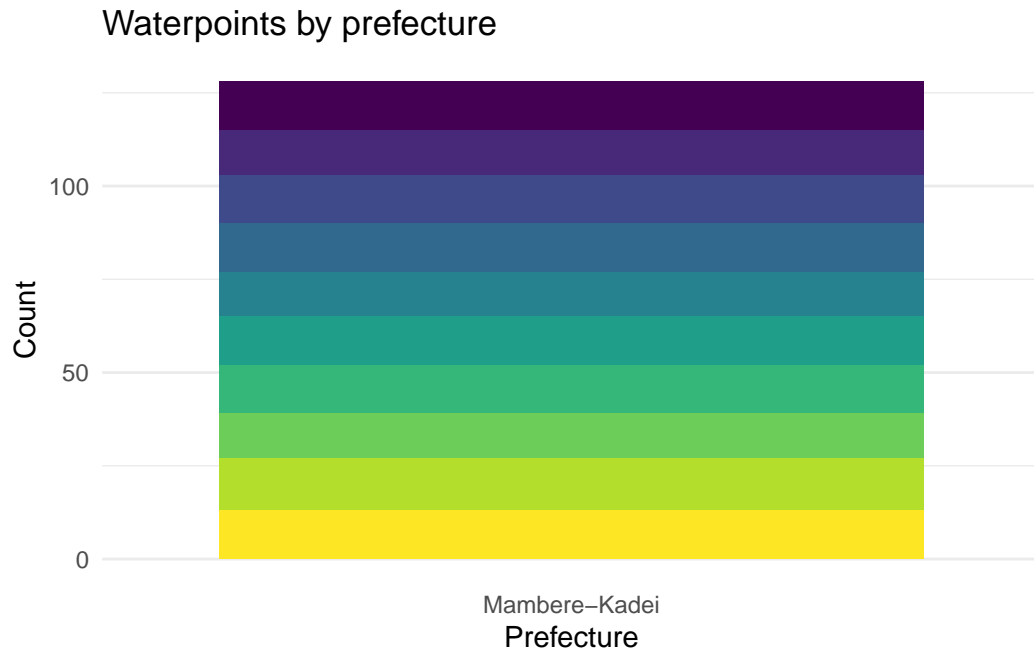
```
 6 Ndembo D          13
 7 Potopoto D        13
 8 Sapoua A          12
 9 Sapoua C - FOSA   14
10 Sosso A           13
```

Both project and baseline have differing waterpoint site name, however both have an even count distribution of wells.

Waterpoints by Prefecture

Waterpoints by prefecture
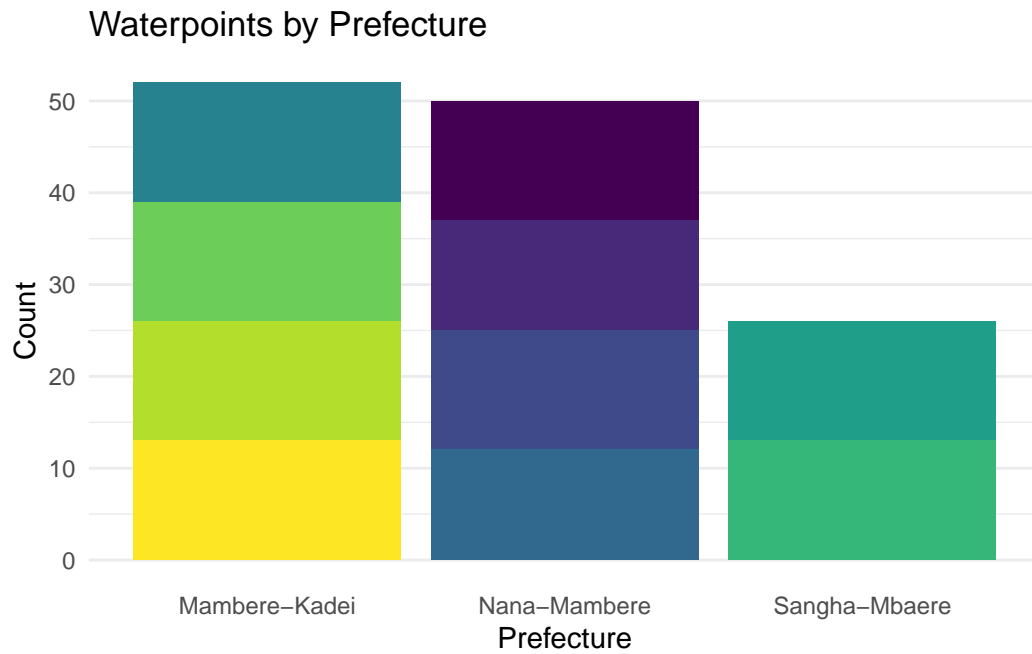
Clearly, this graph visualizes the waterpoint site names at each prefecture in both project and baseline surveys.